The Gini index of speech

THE GINI INDEX OF SPEECH

Scott Rickard and Maurice Fallon

University College Dublin, Dublin, Ireland
{scott.rickard,maurice.fallon}@ucd.ie

ABSTRACT

In which representation is speech most sparse? Time-scale? Time-frequency? Which window generator and length should be used to create the sparsest decomposition? To answer these questions, we propose the Gini index, which is twice the area between the Lorenz curve and the 45 degree line, as a measure of signal sparsity. The Gini index, introduced in 1912, is one of the most common measures of income or wealth distribution and is used to measure the inequity, or sparseness, of wealth distribution. Numerous decompositions of the speech signals in the TIMIT database are used to determine the most sparse standard representation for speech.

1. INTRODUCTION

Sparse signal representations lead to efficient and robust methods for compression, detection, denoising, and signal separation [1, 2]. However, there is no standard practical measure of sparsity. In a strict sense, sparsity means that most signal components are zero. In a practical sense, sparsity means that most signal components are relatively small, and there exists no universal quantitative measurement of this concept.

Sparsity has garnered recent interest in the blind source separation community. In this domain, the goal is, given matrix $\mathbf{Y}$ of the form,

$$\mathbf{Y} = \mathbf{AX} + \mathbf{N} \tag{1}$$

determine matrices $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{N}$ which minimize,

$$\|\mathbf{Y} - \mathbf{AX}\|_F + \lambda \|\mathbf{X}\|_G \tag{2}$$

for matrix cost functions $\|\cdot\|_F$ and $\|\cdot\|_G$ and regularization parameter λ. Often, $\|\mathbf{X}\|_G = \sum_i G(\mathbf{x}_i)$ where $\mathbf{x}_i$ is the ith column of $\mathbf{X}$ and $G(\mathbf{x})$ measures the sparseness of vector $\mathbf{x}$. From the infinite number of possible solutions, we prefer solutions with sparse representations because the original signals themselves have sparse representations.

Often, $G(\mathbf{x})$ is of the form,

$$G(\mathbf{x}) = \sum_{j=0}^{N} g(x_j) \tag{3}$$

where $x_j,\ j = 1, \ldots, N$ are the N components of vector $\mathbf{x}$. Some commonly used $G(\mathbf{x})$, investigated in [2], include:

$$\begin{aligned}
l^0 &: \quad \|\mathbf{x}\|_0 = \#\{j, x_j \neq 0\}/N \\
l^0_\epsilon &: \quad \|\mathbf{x}\|_{0,\epsilon} = \#\{j, |x_j| \geq \epsilon\}/N \\
l^p &: \quad \|\mathbf{x}\|_p = (\sum_j |x_j|^p)^{1/p} \\
\tanh_{a,b} &: \quad \sum_j \tanh\left(|a x_j|^b\right) \\
\log &: \quad \sum_j \log\left(1 + x_j^2\right) \\
u^0_\theta &: \quad \min_{i,j}(x_{(i)} - x_{(j)}) \text{ s.t. } \frac{i-j}{N} \geq \theta \ \& \ x_{(j)} \leq 0 \leq x_{(i)} \\
&\qquad \text{for ordered data, } x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}
\end{aligned}$$

Plots of the individual measures of sparseness listed above are shown in Figure 1.

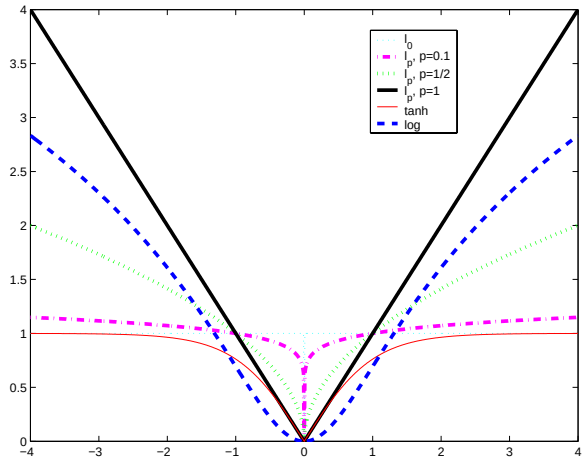

Fig. 1. Component sparsity contributions as a function of component amplitude.

Note, for all these measures, the more sparse the signal, the closer to zero its sparseness measure. For example, the l^0 norm penalizes any non-zero component equally with a contribution of $1/N$ to the sparseness measure where N is

the number of components. l_ϵ^0 is often used when noise is present, as the noise results in very few components being truly zero, despite the fact the representation is still sparse in an intuitive sense. As optimization using l_ϵ^0 is difficult because the gradient yields no information, l^p is often used in its place, with $p < 1$. $\tanh_{a,b}$ is sometimes used in place of $l^p,\ p < 1$, because it is limited to the range $(0, 1)$ and better models l^0 and l_ϵ^0 in this respect. The fact that $l^p,\ p < 1$ and $\tanh_{a,b}$ are concave enforces sparsity. That is, a representation is more sparse if we have one large component, rather than dividing up the large component into two smaller ones. The $\log$ measure is concave outside some range, but convex near the origin, which in effect spreads the small components. The last measure u_θ^0 measures the smallest range around the origin which contains are certain percentage of the data.

All of these measures are somewhat arbitrary and most depend heavily on the choice of parameter settings with the exception of the l^0 norm. The l^0 norm however is not practical in the presence of noise. So we endeavor to find a measure of sparseness which has no parameters and can handle noise. It turns out that economists interested in the study of the distribution of wealth have such a measure, called the Gini index, which we formally define in the next section. For now, we will focus on some desirable properties that a sparse measure should have. The economist Dalton in [3] cited four properties that a sparse measure should satisfy. The four properties are (as in [4]):

- (Dalton's 1st Law) Robin Hood decreases sparsity. Stealing from the rich and giving to the poor, decreases the inequity of wealth distribution (assuming you don't make the rich poor and the poor rich).
- (Dalton's modified 2nd Law) Sparsity is scale invariant. Multiplying wealth by a constant factor does not alter the effective wealth distribution.
- (Dalton's 3rd Law) Adding a constant decreases sparsity. Give everyone a trillion dollars and the small differences in overall wealth are then negligible.
- (Dalton's 4th Law) Sparsity is invariant under cloning. If you have a twin population with identical wealth distribution, the sparsity of wealth in one population is the same for the combination of the two.

We argue that all these principles seem reasonable from a sparse signal representation point of view, and we add two desired properties to this list:

- (Proposal 1) Bill Gates increases sparsity. As one individual becomes infinitely wealthy, the wealth distribution becomes as sparse as possible.
- (Proposal 2) Babies increase sparsity. Adding individuals with zero wealth to a population increases the sparseness of the distribution of wealth.

It can be shown that all are satisfied by the Gini index, and this is not the case for any of the previously discussed measures. All of these six principles can be thought of in terms of components of a representation instead of individuals in a population and component strength (magnitude) in place of individual wealth.

The rest of the paper is as follows. We discuss the formula for the Lorenz curves and the Gini index in Section 2 and perform experiments on speech signals taken from the TIMIT database in Section 3 to determine in which representation is speech most sparse. We present conclusions in Section 4.

2. LORENZ CURVES AND THE GINI INDEX

Given data, $\mathbf{x} = \{x_1, x_2, \ldots x_N\}$, we order the data according to magnitude, $|x_{(1)}| \leq |x_{(2)}| \leq \cdots \leq |x_{(N)}|$. The Lorenz curve, originally defined in [5], is the function with support $(0, 1)$, which is piecewise linear with $N + 1$ points defined,

$$L\left(\frac{i}{N}\right) = \sum_{j=1}^{i} \frac{|x_{(i)}|}{\sum_{k=1}^{N} |x_k|}, \quad \text{for } i = 0, \ldots, N \qquad (4)$$

Note, $L(0) = 0$ and $L(1) = 1$.

The Gini index, originally proposed (in English) in 1921 in [6], is the twice the area between the Lorenz curve and the 45 degree line. The area underneath the Lorenz curve is,

$$A(\mathbf{x}) = \frac{1}{2N} \sum_{n=1}^{N} \left(L\left(\frac{n-1}{N}\right) + L\left(\frac{n}{N}\right) \right) \qquad (5)$$

The Gini index is then simply,

$$G(\mathbf{x}) = 1 - 2A(\mathbf{x}). \qquad (6)$$

Figure 2 shows the Lorenz curve and Gini index for four simple vectors. Note that the distribution in which all individuals have equal wealth is the least sparse and the distribution in which all the wealth is concentrated in one individual is the most sparse.

$G(\mathbf{x})$ has many nice properties:

- A representation with equal wealth distribution has $G(\mathbf{x}) = 0$, no sparsity.
- (Dalton's 1st & 2nd Law) $G(\mathbf{x})$ satisfies the Robin Hood Principle and is scale invariant.
- (Dalton's 3rd Law) $G(\mathbf{x} + k) \to 0$ as scaler $k \to \infty$.

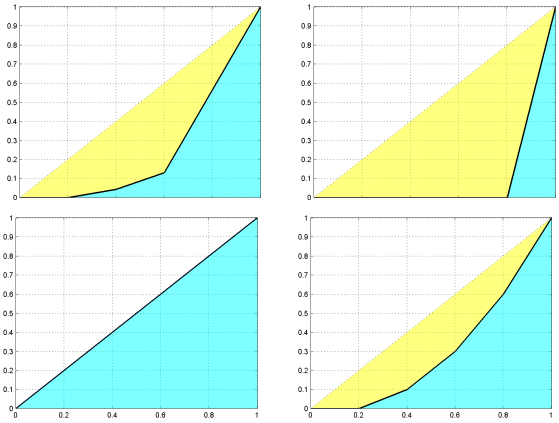

Fig. 2. Lorenz curve for [0 1 2 10 10] (top left), [0 0 0 0 1] (top right), [1 1 1 1 1] (bottom left), and [0 1 2 3 4] (bottom right). The Gini index is twice the lightly shaded/yellow area. The Gini indexes are 0.5043, 0.8, 0.0, and 0.4, respectively.

- (Dalton's 4th Law) $G(\{x_1, x_2, \ldots x_N\})$ is identical to $G(\{x_1, x_1, x_2, x_2, \ldots, x_N, x_N\})$.
- (Proposal 1) As one component of a representation goes to infinity, $G(\mathbf{x}) \rightarrow 1$.
- (Proposal 2) If an infinite number of zero components are added to a vector, $G(\mathbf{x}) \rightarrow 1$.

3. THE GINI INDEX OF SPEECH

In this section we present results of using the Gini index to measure the sparsity in the time-frequency and time-scale domains of thirty speech signals taken from the TIMIT database. For each speech file, the TIMIT annotations were used to determine the starting point of the speech and the silence before that point was removed. Approximately one second of speech ($2^14 = 16384$ samples) was analyzed for each file and each decomposition consisted of 16384 components. That is, adjacent windows were used in the time-frequency case and no extension technique was used for the wavelet decompositions. The sampling rate of the signals was 16kHz and they were transformed into the time-frequency and time-scale domain using various windows and wavelet filters. For each decomposition, the Gini index of the squared magnitude of the components was measured. In each case, the decomposition consisted of Each data point presented in the figures in this section represents the average Gini index over the thirty speech files.

3.1. Time-frequency

For the time-frequency representations, we used the windowed Fourier transform using each windows depicted in Figure 3 for window lengths = $\{2^0, 2^1, \ldots, 2^14\}$. The results of the tests are shown in Figure 4. The figure shows that speech is most sparse in the time-frequency domain when using a Hann, Hamming, or triangular window shape of length approximately 1024 samples, which corresponds to 64 ms. Example Lorenz curves for the time domain and a time-frequency domain representation of a sample speech file are shown in Figure 5.

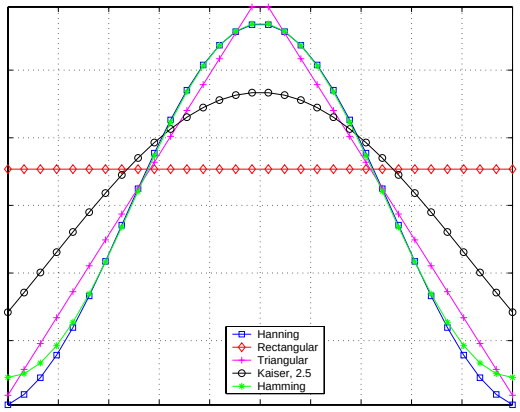

Fig. 3. Window shape comparison.

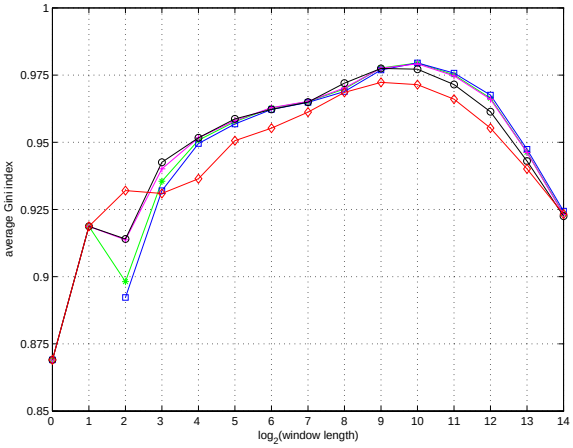

Fig. 4. Gini index as a function of window size for common window types. In the time-frequency domain, speech is most sparse when using a Hann, Hamming, or triangular window shape of length approximately 1024 samples, which corresponds to 64 ms.

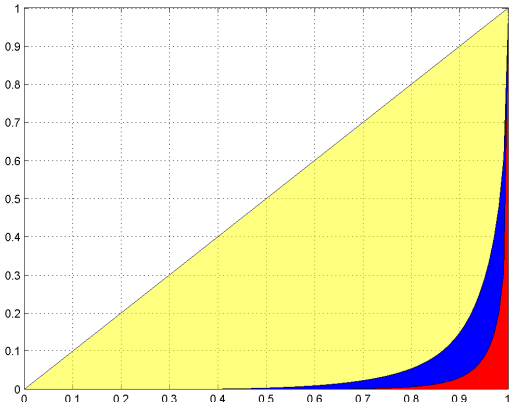

Fig. 5. Lorenz curve for speech in the time domain (darker shade/blue) and time-frequency domain (lighter shade/red). Gini index of 0.90 (time-domain) and 0.97 for TF domain (window length 64 ms).

3.2. Time-scale

In order to compare the time-frequency methods with the time-scale methods, care was taken to ensure that the sparsity measurements for the time-scale methods were not biased by the different data border treatment methods. For example, standard wavelet decomposition methods extend the data with zeros, symmetrically, with zeroth (or first) order extension, or periodically. In order to avoid dependence of the results on extension technique, we chose to instead assume that we had an infinitely long speech signal and were looking at a decomposition snapshot at a given period in time. Thus, components at lower levels of the decomposition depend on data coming from outside the one second analysis window considered in the time-frequency tests. This is an unavoidable consequence when comparing time-frequency and time-scale methods. Either the results will depend on the extension technique, or, the time-scale method must be allowed to see more data than the time-frequency method in order to produce the same number of decomposition coefficients. We chose the latter for this paper, which limits the depth of the decomposition that we can perform because the original signals from the TIMIT database contain under ten seconds of data for each file.

The results of performing the decomposition using Daubechies wavelets of length $2, 4, \ldots, 12$ for various decomposition levels are shown in Figures 6, 7, and 8. A full wavelet packet decomposition, corresponding to a binary tree of depth L, would have $A(L) = A(L-1)^2 + 1$, where $A(1) = 1$, possible lossless representations. This number grows quite rapidly (it is asymptotic to c^{2^n}, $c \approx 1.226$); $A(1) = 1, A(2) = 2, A(3) = 5, A(4) = 26, A(5) = 677, A(6) = 458330, A(7) = 210066388901, \ldots$, see sequence A003095 in [7]. Determining the sparsest binary tree decomposition from the $A(L)$ possible decompositions is thus computationally not feasible. Thus, we measured the Gini index of three simple decompositions (Example decompositions for these three schemes are depicted in Figure 9):

- the standard wavelet decomposition which corresponds to applying the low-pass and high-pass filter to the low-pass output at the previous level,
- the wavelet packet decomposition (all the leaves) at a given level,
- the decomposition arising from a greedy algorithm which starts at the leaves and chooses the sparser of the two leaves or the parent for each pair of leaves at the deepest level, replaces the parent decomposition with the winning result in each case, repeats the process one level higher until reaching the top node.

Figure 6 shows that for the standard wavelet decomposition, the sparsity increases as the filter length and decomposition depth increases. However, it plateaus for level ≥ 4 and filter length ≥ 6 at a Gini index ≈ 0.96, so there is no justification from a sparsity standpoint for using longer filters or more levels of decomposition, as these require additional computation. Figure 7 shows that speech is most sparse at level 8 for all the filter lengths tested when considering all the leaves at a given level. Figure 8 shows that a selective decomposition increases the sparsity until level 8, where it levels off. A comparison of the three composition methods for the Daubechies filter of length 12 is shown in Figure 10. From the figure, we conclude that there is not much to be gained by using the greedy algorithm as the performance of the level method obtains a Gini near the maximum value obtained and is less computationally intensive. Tests on coiflets and symlets resulted in similar conclusions with slightly inferior results.

4. COMPARISON AND CONCLUSIONS

Comparing the results of the time-frequency and time-scale sparsity measurements, we come to the conclusion that speech is slightly more sparse in the time-frequency domain. However, far more important than the choice of domain is the choice of window length (or filter length/level decomposition, in the time-scale case). For appropriate choices, continuous speech is more than 97.5% sparse when using the Gini index as a measure of sparseness.

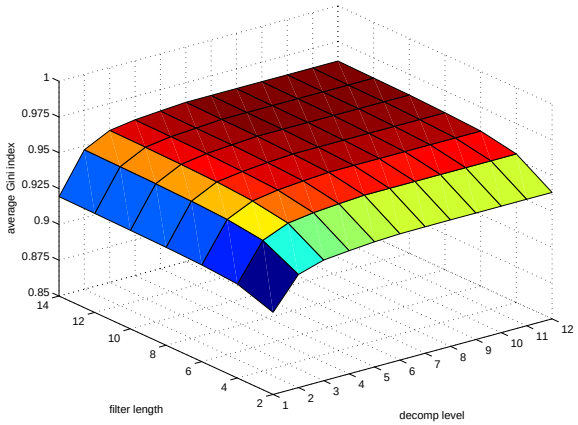

Fig. 6. Sparsity for standard wavelet decomposition using Daubechies wavelets of length $2, 4, 6, 8, 10, 12, 14$ and decomposition level 1 through 12.

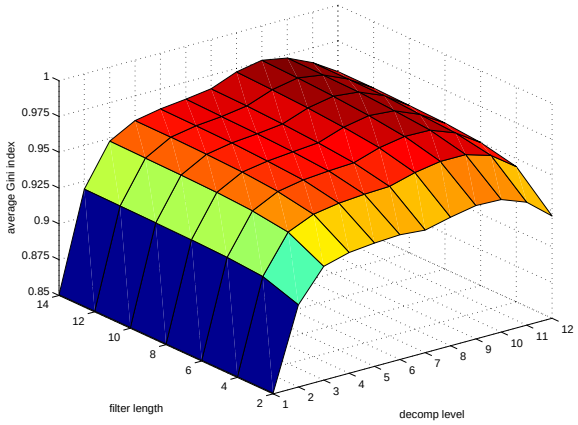

Fig. 7. Sparsity for wavelet packet decomposition using Daubechies wavelets of length $2, 4, 6, 8, 10, 12, 14$ at decomposition level 1 through 12.

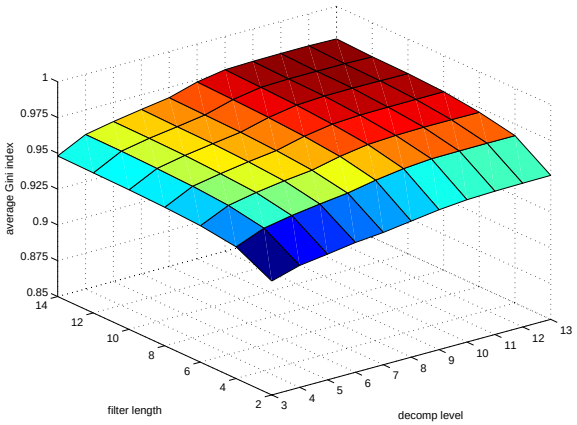

Fig. 8. Sparsity for wavelet packet decomposition using Daubechies wavelets of length $2, 4, 6, 8, 10, 12, 14$ and decomposition level 3 through 13. Binary tree selection using greedy selection method.

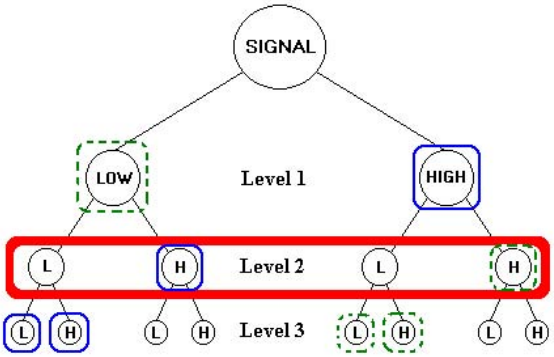

Fig. 9. The simple wavelet packet decompositions. Blue/solid line shows the standard wavelet decomposition for level 3, the red/thick line shows the level decomposition for level 2, and the green/dashed line shows one possible greedy decomposition for level 3.

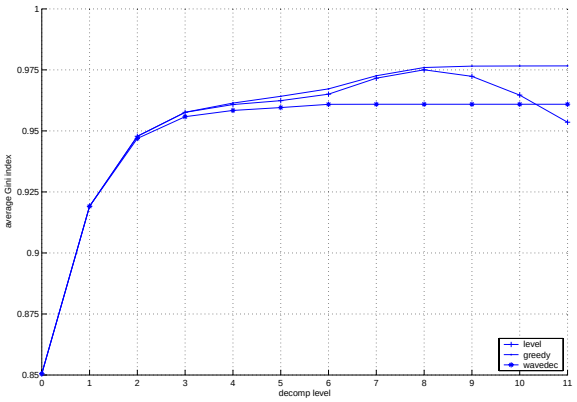

Fig. 10. Comparison of Gini index for decomposition techniques for Daubechies filter length 12.

5. REFERENCES

[1] D. L. Donoho. Sparse components analysis and optimal atomic decompositions. constructive approximation. *Constructive Approximation*, 17:353–382, 2001.

[2] J. Karvanen and A. Cichoki. Measuring sparseness of noisy signals. In *ICA03*, 2003.

[3] H. Dalton. The measurement of the inequity of incomes. *Economic Journal*, 30:348–361, 1920.

[4] B. C. Arnold. *Majorization and the Lorenz Order: A Brief Introduction*. Springer-Verlag, 1986.

[5] M. O. Lorenz. Methods of measuring concentrations of wealth. *J. Amer. Stat. Assoc.*, 1905.

[6] C. Gini. Measurement of inequality of incomes. *Economic Journal*, 31:124–126, 1921.

[7] N. J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences*. published electronically at `http://www.research.att.com/~njas/sequences`, 2004.